Project Summary
SDCI Net: Collaborative Research: An integrated study of datacenter networking and 100 GigE wide-area networking in support of distributed scientific computing

Datacenter and wide-area networking technologies, such as InfiniBand, Fibre Channel, and Ethernet, have seen significant advances, transport protocols have been enhanced, parallel file systems have improved disk access throughput, and applications for parallel transfers of large files have been developed. Nevertheless, scientists still see high variability in end-to-end application-level performance, and often experience low throughput or increased latency.

In collaboration with the Community Earth System Model (CESM) project, this proposal seeks to determine the reasons for these end-to-end application level performance problems by gathering and analyzing various measurements as CESM scientists run their applications as part of their daily research. The knowledge gained will be used to create realistic configurations of parallelized Message Passing Interface (MPI) applications and file-transfer applications on a controlled experimental testbed that consists of two datacenters, located at National Energy Research Scientific Computing Center (NERSC) and Argonne National Laboratory (ANL), interconnected by an ESnet-run 100 GigE wide-area Advanced Networking Initiative (ANI) prototype network (see letters of support from NERSC and ESnet). Three sets of experiments will be run. The first set consists of CESM MPI applications and a set of benchmark programs that will be run within the NERSC datacenter. The second set will study the feasibility of using wide-area GPFS for MPI I/O. The third set will evaluate two or more file-transfer applications popular in the science community, such as GridFTP and bbFTP, to obtain wide-area throughput as close to 100 Gbps as possible.

The emergent Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) technology is interesting as it can be combined with rate-guaranteed (dedicated) Ethernet VLAN based virtual circuits, and it uses the InfiniBand transport protocol rather than TCP. This integrated datacenter/wide-area networking alternative (RoCE with Ethernet VLANs, in which there is no TCP/IP) will be evaluated and compared with a solution that uses Internet Wide Area RDMA Protocol (iWARP) inside the datacenter and IP routed wide-area paths (in which TCP is the transport protocol).

The **intellectual merit** of the proposed project lies in its systematic scientific approach consisting of the following steps:

1. determine the reasons for poor end-to-end application-level performance
2. use this knowledge to upgrade the weakest components by experimenting with multiple parallelized MPI applications and commonly used file-transfer applications on a variety of datacenter and wide-area networking configurations created in a controlled environment, and
3. transfer these upgraded applications and networking solutions to CESM scientists for field testing, make modifications based on their experience, and then transfer the developed technology to other science communities.

The **broader impact** component of the proposed project consists of three sets of activities:

1. integration of education through the creation of a course module on datacenter networking, and the involvement of undergraduate students in this research at all three institutions, UVA, UNH and NCAR
2. participation in diversity enhancing programs such as high school visitation weekends for female and African American students, and summer outreach programs, organized through centers for diversity
3. active promotion of the modified applications and networking solutions created by this project among the CESM scientists as well as scientists from six scientific areas as organized by DOE's Office of Science.

SDCI Net: Collaborative Research: An integrated study of datacenter networking and 100 GigE wide-area networking in support of distributed scientific computing

1 Introduction

1.1 Problem statement

Datacenter networking technologies include InfiniBand for Inter-Processor Communications (IPC), Fibre Channel for storage, and Ethernet for wide-area network access. InfiniBand's excellent low-latency performance makes it well suited for parallelized scientific Message Passing Interface (MPI) applications that require low latency communications. Parallel file systems such as General Parallel File System (GPFS) and multi-Gb/s Fibre Channel have improved disk access performance.

Advances in wide-area networking include increased transmission speeds (100 GigE), as well as the availability of high-speed rate-guaranteed dynamic virtual circuit service on core research-and-education networks, such as ESnet, Internet2, and NLR. Improved transport protocols for IP-routed paths include TCP based approaches such as [1–5] and UDP based approaches such as [6–8], and new transport protocols for virtual circuits include [9–11]. Applications such as GridFTP [12] and bbFTP [13] have led to improvements in file transfer throughput through the use of parallel streams and other techniques.

In spite of all these advances, scientists still see room for improvement in end-to-end application level performance. The National Center for Atmospheric Research (NCAR) PI on this proposed project, as a member of Community Earth System Model (CESM) project, has first-hand experience with both wide-area GridFTP and wide-area GPFS. As part of a NSF funded PetaApps project, this PI transferred about 120 TB of output data over a 4 month period from National Institute of Computational Sciences (NICS) to NCAR using GridFTP at about 70 to 200 MBytes/sec across Teragrid (with 10 Gb/s connectivity). Significant variability in throughput was observed on these transfers. Furthermore, other university participants in the CESM community experience much lower application-level transfer rates. Transfer rates obtained while using the wide-area GPFS filesystem provided by San Diego Supercomputer Center (SDSC) are also dissatisfying and exhibit even larger variability. Fundamentally, it is still not possible for end users to utilize the full potential of high-performance wide-area networking.

Therefore, the *first problem* addressed in this work is to gain an understanding of the reasons for these end-to-end application level performance problems. A data collection/analysis activity is planned to obtain logs and other measurements from CESM scientists' applications as they carry out their daily research.

The knowledge gained from the analysis activity will be used to create realistic configurations of parallelized Message Passing Interface (MPI) applications and file-transfer applications on a controlled experimental testbed that consists of two datacenters, located at National Energy Research Scientific Computing Center (NERSC) and Argonne National Laboratory (ANL), interconnected by an ESnet-run 100 GigE wide-area Advanced Networking Initiative (ANI) prototype network (see letters of support from NERSC and ESnet). Thus the *second problem* addressed in this work, through this controlled experimentation activity, is to evaluate different datacenter and wide-area networking solutions to identify the configurations under which the best performance can be achieved for scientific MPI applications and large file transfers.

Among the datacenter networking solutions is an emergent technology called Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) [14]. This technology is interesting for three reasons:

- It uses RDMA, which enables direct memory to memory transfers of data without operating system involvement. This reduces latency and improves throughput for the network I/O operations required within the computers. The GPFS RDMA code developed for InfiniBand can now be used across Ethernet networks, and since Ethernet networks offer wide-area connectivity, wide area GPFS with RDMA can be explored.
- RoCE uses the InfiniBand transport protocol and not TCP. As Ethernet networks provide wide-area

connectivity, the suitability of InfiniBand transport protocol across wide-area paths can be studied. For example, how does the congestion control algorithm in this transport protocol perform on high bandwidth-delay product paths?

- There is an opportunity to integrate RoCE with wide-area Ethernet VLAN based virtual circuits. The protocol stack in this scenario has no TCP/IP and is yet usable across wide-area paths. The protocol stack is InfiniBand transport and network layers running above Ethernet, and all en route switches forward Ethernet frames with only MAC address and VLAN ID lookups.

An alternative approach that combines RDMA with Ethernet is called Internet Wide Area RDMA Protocol (iWARP) [15]. This solution enjoys the advantages of RDMA while using the TCP/IP stack over Ethernet. Since Ethernet VLAN based virtual circuit service is not available universally, while IP-routed datagram service is available, the iWARP solution is easier to deploy. Features of the RoCE and iWARP RDMA over Ethernet solutions are summarized in Table 1. Both solutions are implemented in 10 Gb/s Ethernet network interface cards, and are referred to as RoCE RDMA Network Interface Cards (RNICs) and iWARP RNICs.

Table 1. RDMA over Ethernet solutions

Solution	Protocols	Corresponding wide-area networking solution
RoCE	RDMA over InfiniBand transport protocol over Ethernet	Ethernet VLAN based virtual circuit
iWARP	RDMA over TCP/IP over Ethernet	IP-routed (datagram) path

The *third problem* addressed in this work is the issue of new technology adoption by scientists in the field. While we will use a subset of the CESM community for initial testing of our technology, we plan to extend our technology transfer effort to other science communities.

1.2 Solution approach

Four steps are planned.

1.2.1 Data analysis to gain an understanding of CESM projects' wide-area networking needs

The goal of this work activity is to determine the exact source of the variability in file transfer throughput, poor average transfer rates for universities, and poor performance of wide-area GPFS for the CESM project. Various monitoring mechanisms will be used to collect data as scientists run applications for their research usage.

1.2.2 MPI applications

The goal of this step is to evaluate different networking options for parallelized MPI applications. One or more MPI applications, instrumented with the Integrated Performance Monitoring (IPM) application profiling technique [16], will be executed within a computer cluster located at NERSC (see attached letter of support), which can be configured with three IPC interconnects: InfiniBand, RoCE and iWARP. Measurements collected will be analyzed for application-level performance metrics such as Sustained System Performance (SSP) [17], as well as lower level metrics such as communication latency and throughput.

While the above experiments compare intra-cluster interconnects, this next set of experiments are across wide-area paths. Given the low latency requirements of parallelized scientific programs, wide-area networking is not proposed here for communications between the parallel processes. Instead it is only proposed for MPI I/O calls. The question is whether wide-area latency can be tolerated if the file systems being accessed by MPI I/O calls are located remotely. The advantage to scientists of this solution is that

they can store their data locally while using the compute power of remote clusters, or store their data on a different cluster, e.g., such as IU's Data Capacitor while using NERSC's compute facilities to run their scientific codes. These experiments will compare the two solutions shown in Table 1, i.e., RoCE with wide-area VLAN virtual circuits versus iWARP with wide-area IP-routed paths. The 100 GigE ANI prototype network supports both IP-routed and VLAN configurations. Since the NERSC and ANL clusters implement GPFS, this is the only file system that will be used.

1.2.3 Large file transfers

There are a number of projects that require the data movement of large (TB to PB) data sets, as listed in the reports resulting from network requirements workshops run by ESnet [18]. The goal of this step is to determine how best to leverage these datacenter and wide-area networking technologies to increase end-to-end application level file-transfer throughput to 100 Gbps.

File transfer applications typically employ socket based APIs rather than MPI. Our University of New Hampshire (UNH) PI has implemented an extended-sockets API library (called EXS) utilizing the Open Fabrics Enterprise Development (OFED) library that enables the use of RDMA NICs (RNICs). These two libraries enable the use of RNICs for any applications that use the TCP socket API. Two or more commonly used file transfer applications, such as bbFTP, bbcp, GridFTP, scp, and sftp, will be modified to use these two libraries thus enabling the use of RDMA and high-speed wide-area networking technologies for increased file transfer throughput.

To enable the use of high-speed Ethernet VLAN virtual circuits, an Inter Domain Controller (IDC) interface module will also be integrated with the file transfer application. This module was developed at UVA starting with a Java client provided by the ESnet On-demand Secure Circuits and Advance Reservation System (OSCARS) project [19], which developed the protocols and software for dynamic circuit scheduling, provisioning and release.

The modified file-transfer applications will be tested on the NERSC and ANL clusters across the 100 GigE ANI prototype network. The goal is to demonstrate 100 Gb/s throughput at the end-to-end application level. End-to-end application level throughput will be measured with both the solutions shown in Table 1, i.e., RoCE with wide-area VLAN virtual circuits, and iWARP with wide-area IP-routed paths.

1.2.4 Trials and technology transfer

The goal of this step is to initiate the adoption of networking solutions and modified applications developed in this project by scientists. We plan to run a series of trials between NERSC, ANL, NCAR, NICS, and SDSC which represent either sources or sinks of CESM generated data. All these organizations are connected to ESnet and since ESnet offers both IP-routed and virtual circuit service, both options can be used. Experiments will be planned based on the availability of InfiniBand, iWARP and RoCE NICs on the clusters. In addition, we plan to execute a concerted effort toward achieving technology transfer to other scientific research teams in areas such as high-energy physics, earth systems grid, etc.

1.3 Anticipated project outcomes

Our anticipated project outcomes includes the public release of our open-source software as listed in Table 2, and a set of publications and reports as listed in Table 3. We will initially deploy and test our software within the most data-intensive subgroup of the CESM user community. This group can immediately benefit from the enhanced capability and will likely provide valuable feedback regarding usability. Based on initial feedback we intend to integrate our enhanced solutions directly into the CESM workflow which would benefit the entire CESM user community. As many scientific disciplines have the need to efficiently utilize high-performance networking, we anticipate wide adoption of our developed software and solutions.

These project deliverables will be disseminated with targeted emails to, and follow-up discussions with,

Table 2. Modified applications and open-source software libraries

File transfer applications such as bbFTP, bbcp, GridFTP, scp and sftp	Modified to incorporate EXS library and to interface with IDC client for virtual circuits
EXS library	An extended-sockets API to interface socket based applications with RDMA hardware

Table 3. Publications and reports

Data analysis for CESM community transfers and WAN GFPS	Data will be collected at several CESM participating institutions
Intra-datacenter CESM and other MPI application experiments	NERSC cluster with InfiniBand, RoCE and iWARP
Inter-datacenter wide-area GPFS for MPI I/O	Tests between NERSC and ANL on ANI prototype network
Inter-datacenter file transfers	100 Gb/s throughput tests between NERSC and ANL on ANI prototype network
Trials of the MPI I/O and file transfer applications	Between NCAR and other datacenters for CESM scientific applications
Report on broader-impact activities	Measurable value from our education and diversity integration activities

specific individuals in other scientific research projects. For example, we expect our intra-datacenter experimental results to be of particular interest to datacenter infrastructure planners. Inter-datacenter wide-area experimental results will be presented to scientific application developers and users. Travel funds have been allocated for participation in meetings and conferences, where the project team members will have an opportunity to meet in person with scientists, application developers, and infrastructure providers. Our report on broader-impact activities will be disseminated to heads of diversity centers, as well as Deans of Engineering schools.

1.4 Division of roles in this multi-institution collaboration

The University of Virginia (UVA) PI is an expert in wide-area networking technologies such as Ethernet, VLAN, MPLS, GMPLS, SONET, WDM, dynamic circuit networking, transport protocols, and hybrid networking. The UNH PI is an expert in datacenter networking technologies such as RDMA, iWARP, Converged Ethernet (CE), RoCE, OFED software, and Extended Sockets (ES) API. The NCAR PI is a leader within the CESM community and an expert on optimizing MPI communication algorithms. As a Gordon Bell winning computational scientist, he is a skilled and vocal consumer of cyber-infrastructure and directly involved in very large TRAC and PRAC allocations. The NCAR senior person is an expert in science gateway development and associated grid middleware. A detailed breakdown of work activities and primary responsibilities for final deliverables are provided in Section 6.

1.5 Management plan

The project is tightly integrated and requires close interaction as seen in the project timeline provided in Section 6. Therefore, the three groups, at UVA, UNH and NCAR will meet regularly via weekly/biweekly conference calls. In addition, travel funds have been allocated for yearly trips to each other's location both for the PI/co-PI and students. A Wiki web site will be created at UVA for sharing documents and

collaborative space for use by the whole team. In addition, a project web site, as well as the NMI Build and Test Facility, will be used for public dissemination of newly developed software programs and technical reports. Funds have been allocated to travel to Sunnyvale to meet with NERSC and ESnet personnel.

2 Background

This section provides background material on high-speed datacenter networking technologies, high-speed wide-area networks, and the scientific research project, Community Earth System Models (CESM).

2.1 Datacenter networking

To reduce CPU load and latency, TCP Offload Engines (TOE), which are hardware implementations of TCP/IP on the NICs, eliminate one copy from application memory to kernel memory. RDMA reduces this even further to a "zero-copy" solution. This section reviews RDMA, InfiniBand, iWARP, RoCE, OFED software and ES-API.

RDMA The RDMA Protocol (RDMAP) [20] consists of four variants of a `Send` operation, all of which are writes to untagged buffers at the data sink that have not been explicitly advertised before, an `RDMA Write` operation to a previously advertised buffer at the data sink, an `RDMA Read` operation that is used to transfer data to a tagged buffer at the data sink from a tagged previously advertised buffer at the data source, and a `Terminate` operation. An `RDMA Send` operation is used by the upper layer (user of RDMAP) to send a short advertisement identifying the location of the tagged buffer and size of the data to be transferred, allowing the data sink to subsequently issue a `Read Request` message to "pull down" the data. Conversely, the upper layer on the receive size could use an `RDMA Send` operation to specify a tagged buffer location and size to its counterpart at the data source, allowing the latter to subsequently execute an `RDMA Write` operation to send the data. The RDMA `Message Size` field is 32 bits allowing for the transfer of messages up to 4 TB in length.

InfiniBand This technology is a low-latency high-performance computer cluster interconnect utilizing a switched fabric [21]. The InfiniBand Architecture (IBA) is packet-switched network with its own set of transport, network, data-link and physical layers. It operates with an MTU of 4096 bytes, and uses a 16-bit Local Identifier (LID) in the packet header for intra-subnet communications. The transport layer supports five types of services: reliable connection, unreliable connection, reliable datagram, unreliable datagram, and raw datagram (which does not use IBA transport layer). RDMA read and RDMA write operations are allowed to only use the reliable connection or reliable datagram transport services, while the RDMA send operation can use any of the transport services except raw datagram. Queue pairs (QP) are used at the sender and receiver. The destination QP is identified in the IBA transport header, comparable to the destination port number used in a TCP header. For reliable service the IBA transport header carries a 24-bit packet sequence number and uses an acknowledgment/time-out ARQ scheme, much like TCP. Thus data is delivered in-sequence and without duplicates. Reliable connection service includes a credit-based flow control mechanism in which the receiver sends the number of available credits (number of messages) in a 5-bit field in the ACK extended transport header (AETH). This again is comparable to TCP's window-based flow control scheme. Maximum message size is limited to 2^{31} bytes.

An InfiniBand Host Channel Adapter (HCA), which is the name used for an InfiniBand NIC, executes RDMA, and the IBA protocol stack (transport, network, link and physical layers). InfiniBand's reliable services allows for the use of RDMA directly on the IBA protocol stack.

iWARP To exploit the zero-copy low-latency RDMA mechanism on low-cost Ethernet, additional protocol layers are necessary because RDMA requires the underlying network to be reliable, and Ethernet is not intrinsically reliable. Unlike InfiniBand's reliable transport services, Ethernet does not have its own transport protocol layer, and relies instead on TCP for reliability. This leads to another mismatch, between

TCP's stream-oriented nature versus RDMA's message-oriented structure. To deal with these issues, two new protocols, Direct Data Placement (DDP) [22], and Marker PDU Aligned (MPA) framing [23], are placed between RDMAP and TCP/IP. Thus iWARP consists of these three protocols: RDMAP, DDP and MPA. iWARP RNICs implement all of these layers: RDMAP, DDP, MPA, TCP, IP and Ethernet in hardware.

RoCE Towards making Ethernet reliable, under the term "Data Center Bridging (DCB)" or "Converged Ethernet," the IEEE has defined standards to support priority-based flow control (IEEE 802.1Qbb) [24], enhanced transmission selection (IEEE 802.1Qaz) [25] to assign bandwidth to 802.1p classes, congestion notification (IEEE 802.1Qau), DCB exchange capabilities (IEEE 802.1AB) [26], and Layer 2 multipathing (IEEE 8021.Qat). The above efforts to make Ethernet reliable have led to solutions such as RDMA over CE (RoCE) [14], in which the InfiniBand transport layer runs on top of reliable Ethernet.

OFED Software, and ES-API The OpenFabrics Alliance (OFA) [27] has organized a software development effort called the OFED [28] to provide open-source software to interface with all types of RDMA interfaces: Infiniband, RoCE and iWARP. This software has an active development community and is now included in Linux kernel distributions.

The OFED software is intentionally not designed as a conventional Application Programming Interface (API). Rather, it is a set of functions, called "verbs", and data structures that can be utilized to build various APIs. This has already been done to produce several versions of the Internet Small Computer System Interface (iSCSI) [29–32] and the Message Passing Interface (MPI) [33, 34].

The use of these verbs and data structures require knowledge of the internal workings of RDMA hardware and its interactions with the OFED software, making it difficult for programmers familiar with conventional sockets programming. It therefore seems desirable to provide a more familiar, general-purpose API on top of the OFED verbs layer to make it easier to write general network programs that can take advantage of the the RDMA hardware. This general-purpose API, Extended Sockets (ES-API), would coexist with more specialized APIs, such as MPI, as illustrated in Fig. 1.

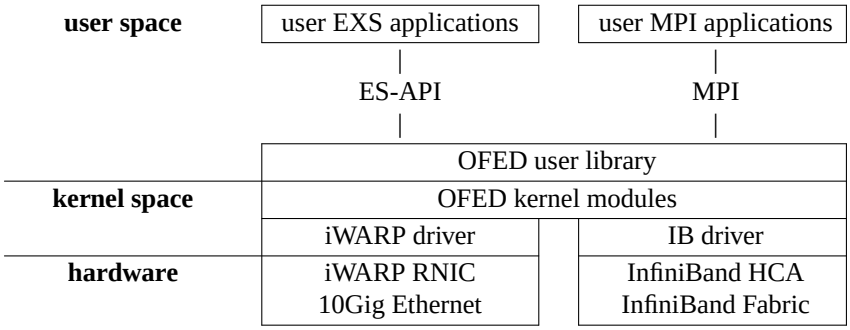

Figure 1. Layering of APIs, OFED, iWARP and InfiniBand

The Extended Sockets API (ES-API) [35] is a specification published by the Open Group [36] that defines extensions to the traditional socket API in order to, among other features, support "zero-copy transmit and receive operations and improved buffer management [37]." It is designed for use in a threaded environment, and contains two major new features: asynchronous I/O and memory registration. These extensions provide a useful method for efficient, high-level access to the OFED stack and RDMA.

2.2 High-speed wide-area network services

There have been significant efforts on new transport protocols to increase throughput for large file transfers across IP-routed networks. Examples of TCP-based approaches include BIC TCP [1], CUBIC

TCP [2], High-speed TCP [3], FAST TCP [4], and Scalable TCP [5], and examples of UDP based approaches include SABUL [6], UDT [7], and RAPID+ [8].

A different problem arose that led to the deployment of virtual-circuit networks. As facilities such as the Large Hadron Collider (LHC) came online, and a number of scientific research projects started moving increasingly large sized data sets, these very large data transfers (TB to PB) caused IP-routed traffic from ordinary applications such as Web browsing, email, and video streaming to experience degraded performance. For example, a high-energy physics project (CMS) large file transfer between Fermilab and the University of Nebraska impacted a streamed video flow being viewed by the University football coaching staff. TCP's congestion control algorithms allow "elephant" (large-sized) flows to consume large portions of the network bandwidth when compared to "mice" flows, as documented in papers such as "The war between elephants and mice" [38]. This led to core research and education network providers such as DOE's ESnet, Internet2, and others, to offer a rate-guaranteed virtual circuit service as a complement to their IP-routed datagram service, and to redirect these large transfer flows to the virtual circuit service thus isolating them from negatively impacting mice flows.

To support rate-guaranteed virtual circuits, ESnet developed a scheduler called OSCARS (On-demand Secure Circuits and Advance Reservation System), which accepts requests for reservations and initiates the provisioning of the virtual circuits at their scheduled start times. When the reserved time interval ends, the circuit[1] resources are released for use by others. This process of admission control coupled with rate policing of all admitted data flows guarantees that there will be no congestion in switch buffers.

ESnet [39], Internet2 [40], Europe's GEANT2 [41], and Canada's Canarie [42] jointly developed an Inter-Domain Controller Protocol (IDCP) [43] to support inter-domain circuits. IDCP consists of message exchanges between domains to support the scheduling, provisioning and release of inter-domain circuits. The *createReservation* message is used for circuit scheduling, and *createPath* message for circuit provisioning. The IDCs communicate with domain controllers and/or the circuit switches located within the networks to provision the circuit. At the scheduled *endTime*, a *teardownPath* message is issued to release the circuit.

Not all site networks that connect to ESnet, such as Fermilab, ANL, LBNL, or regional networks that connect to Internet2 or NLR have deployed virtual circuit networks. Therefore, while some projects such as Terapaths [44] and CHEETAH [45] focus on end-to-end circuits, other projects, such as Lambdastation [46], have developed techniques and software for redirecting flows at customer- or ESnet provider-edge routers on to virtual circuits that just span the backbone network.

2.3 Community Earth System Models (CESM)

The Community Earth System Model (CESM) [47] is an important and widely used community application for climate research. After decades of development by staff at NCAR, the Department of Energy (DOE), and the university community, CESM has matured into a true, multi-component Earth System model with dynamic and fully interacting atmosphere, land, ocean, and sea-ice components. CESM has an impressive resume of scientific achievement, including contributing to the Nobel Prize-winning Fourth Assessment report by the Intergovernmental Panel on Climate Change (IPCC AR-4) [48]. The CESM user base includes a wide variety of users that may utilize computing systems that vary from a laptop to the largest Petascale system currently deployed. We consider two different types of CESM users for this project. Many CESM users perform their computing on small departmental clusters that may consists of 8 to 32 compute nodes. While this group of users could potentially benefit from commercial Web based Cloud services, initial performance results from the Community Atmosphere Model (CAM), the atmospheric component of CESM, are not encouraging [17]. A factor of 11x slowdown for the Amazon EC2 cluster versus a Cray XT4 is observed due to increases in the communication cost. Other users of CESM

[1]The terms "circuits" and "virtual circuits" are used interchangeably in this document.

utilize very large scale parallelism to perform ultra-high-resolution simulations that generate very large amounts of data, e.g., the PetaApps project mentioned in Section 1. The Athena Project [49], a multi-nation project utilized the Athena supercomputer at NICS to generate 900 TBytes of data over the course of 6 months. Other climate projects are currently underway at Oak Ridge National Laboratory and the Geophysics Fluid Dynamics Laboratory that are similar in scope and size. Our NCAR PI is also part of a team that received at PRAC award to perform very large scale climate simulations on the NSF Blue Waters at NCSA. The NSF Blue Waters PRAC will likely generate 6 PBytes of output data.

While it is now possible to generate data at these very large scales, analyzing the output at this scale is challenging for a number of reasons. In the case of the PetaApps project, all 120 TB of output data was transferred back from NICS to NCAR. This decision to transfer output data from the project was made because a suitable data analysis like the "Nautilus" system at NICS did not yet exist at the start of the project. While a tuned version of gridFTP was used for the data transfer, we were only able to achieve a 70 to 200 MB/sec transfer rate, which is a fraction of the available bandwidth between NICS and NCAR. Further the transfer capability was non trivial to configure and use. While transfer of the entire 120 TB PetaApps dataset is feasible given its "modest size", the much larger PByte size datasets will likely never be moved from the computing center where they were generated without a significant improvement in the ability to utilize high-performance wide-area networks. The current lack of mobility makes intercomparisons between the large datasets problematic.

3 Experimental setup

Through a DOE project on scientific cloud computing called Magellan [50], NERSC and ANL have deployed two clusters. These clusters will be interconnected via the DOE-funded Advanced Network Initiative (ANI) wide-area 100 GigE prototype network [51]. Each Magellan cluster consists of compute nodes, I/O nodes for storage, and network nodes, as shown in Fig. 2, and Fig. 3 shows the wide-area 100 GigE ANI network. The Magellan NERSC and ANL clusters connect to the routers at these sites, respectively.

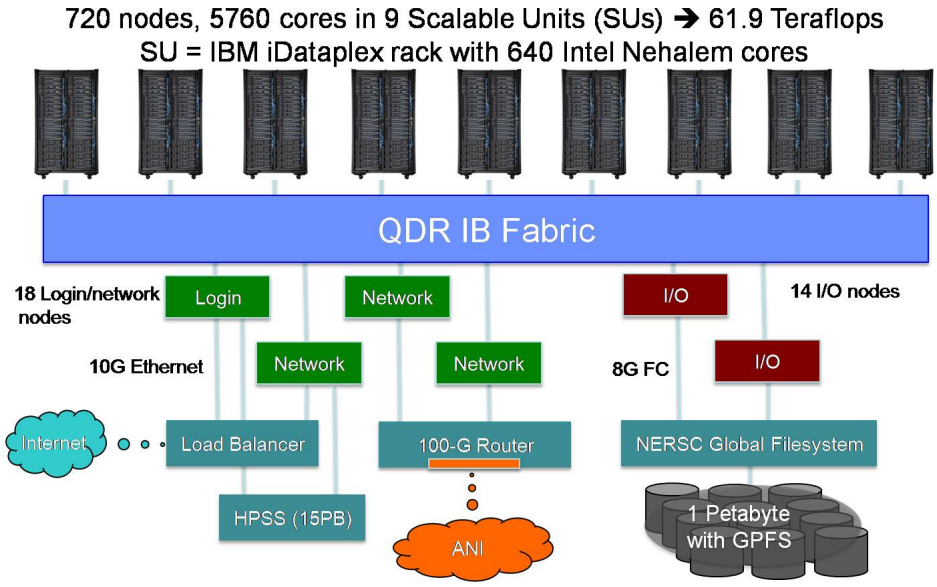

Figure 2. Magellan cluster at NERSC [52]

As shown in Figure 2, the 720 compute nodes of the NERSC cluster have quad-core 2.67GHz Intel Xeon (Nehalem) processors, 24GB memory, and a Mellanox ConnectX VPI PCIe 2.0 5GT/s IB QDR/10GigE

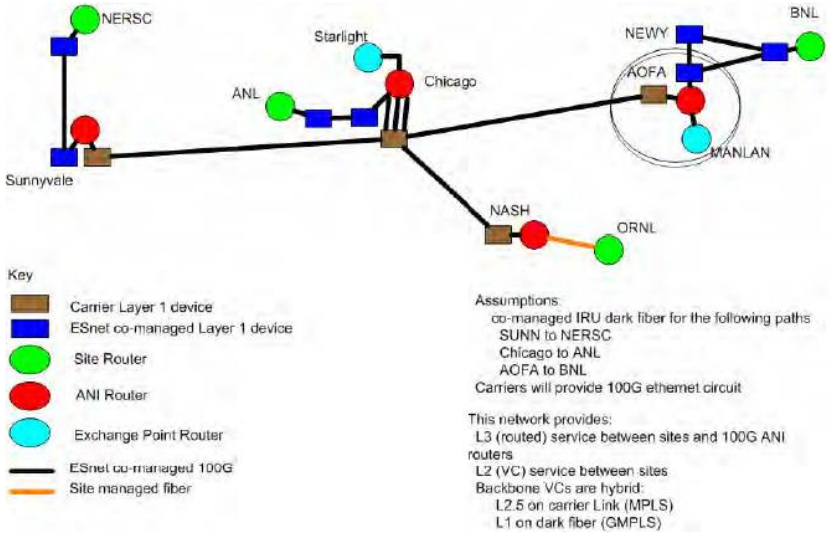

Figure 3. ANI network [51]

interconnect card. The login/network nodes have the same processor, 48GB RAM, the same Mellanox card, and a 10 GigE Chelsio S310E-SR+ card. The I/O nodes have the same configuration as the compute nodes. In addition they have dual-port 8 Gb/s Qlogic Fibre Channel adapters. Login access will be provided to the login/network nodes and compute nodes for our experimental research. File systems will be mounted on the network and compute nodes to access files stored in the GPFS I/O nodes.

4 Proposed work

To achieve the project objectives, our approach consists of four sets of activities, as listed in Section 1.2. All newly developed and modified software will be managed with the NMI Build and Test facility at http://nmi.cs.wisc.edu. All experimental data collected will also be stored in this facility as described in the attached Data Management Plan.

4.1 Data analysis to gain understanding of CESM projects' wide-area networking needs

As noted in Section 1.2.1, CESM scientists' applications and hosts will be instrumented with various monitoring mechanisms, such as GridFTP logging, tcpdumps, IPM profiling for MPI I/O, and PerfSONAR One-Way Ping (OWAMP) [53] delay measurements will be obtained on overlapping paths. Automation scripts will be implemented so that as scientists run applications for their research usage, data is collected and moved to the University of Virginia for analysis. Using Perl scripts and statistical R programs, we will process data from disparate sources to identify the causes of poor performance. As shown in Section 6, this instrumentation and data collection will first be implemented for CESM scientists located at NCAR. But, because CESM is used by a broad and geographically distributed user community, we will also collect and analyze measurements from CESM researchers at other organizations such as U. Miami, George Mason U., ORNL, and LLNL.

4.2 MPI applications

As described in Section 3, the NERSC computer cluster has the Mellanox ConnectX VPI cards in which the ports can be configured as either InfiniBand or 10 GigE. These cards thus support both InfiniBand and RoCE. In addition the 10 GigE Chelsio S310E-SR+ card supports iWARP. Thus, these three datacenter interconnect technologies can be configured on the NERSC cluster to evaluate the performance of several

MPI applications.

While we will focus on CESM, which contains five separate component models each with different MPI communication patterns, we will also augment our analysis with other MPI-based benchmarks. Select applications from the NAS parallel [54], NERSC-6 procurement [55], and HPCC [56] benchmarks will be used to accurately characterize the diversity of high-performance computing message passing traffic [57, 58]. Each benchmark will be instrumented with the IPM application profiling technique.

A subset of MPI applications with MPI I/O disk read/write operations will be selected for the wide-area GPFS experiments between NERSC and ANL across the ANI wide-area network. Again, extensive monitoring tools will be used to collect measurements for analysis of the performance of wide-area GPFS. As described in Section 1.2.2, the application-level performance will be compared for the two options shown in Table 1, i.e., RoCE with wide-area VLAN virtual circuits versus iWARP with wide-area IP-routed paths.

4.3 Large file transfers

As noted in Section 1.2, because file transfer applications use the socket API, the first step is to complete our implementation of the Extended Sockets API, and then upgrade one of more file transfer applications to use this API. Also, sofware development is required to have the file transfer applications invoke the setup of virtual circuits for the RoCE-VLAN experiments.

4.3.1 EXS implementation

The University of New Hampshire Extended Sockets (UNH-EXS) library is designed as a multi-threaded implementation of the ES-API that also provides some additional API facilities [59]. It is an extension of the standard sockets interface that enables applications to transparently utilize all RDMA technologies. It is implemented entirely in user space as a light-weight layer between application programs and the OFED verbs, as shown in Figure 1. Working entirely in user space rather than kernel space makes it much easier to implement and debug software, and it increases the portability of the resulting code.

Although UNH-EXS is still under development, early benchmarks have demonstrated that high-bandwidth utilization and low CPU overhead can be realized in practice [60]. This development of UNH-EXS needs to be completed. It involves adding a number of functions defined in the EXS standard [35] but missing from the current partial implementation, making the error handling more robust, performing more exhaustive testing of combinations of features, writing user-level documentation, and distributing the project to the public.

The iWARP protocol stack can be implemented completely in software for use on end-points that have only Ethernet NICs, not RNICs. Such **softiwarp** end-points will not themselves see any reduction in CPU utilization, but the RNIC at the communicating end-point will. Since most systems have an "overloaded" or "load-sensitive" component, this makes it possible to reduce that component's CPU load by configuring it with an RNIC, and to utilize softiwarp on an ordinary NIC at the other end of the connection. Components implemented on ordinary low-cost platforms can thereby communicate with high-performance clusters without requiring any special coding on the cluster, and without degrading the performance of the cluster. It also provides a path to upgrade systems without RDMA hardware by first installing and testing software using only softiwarp, and then later installing the hardware RNICs. Finally, it means that high-performance clusters can export all their results using only iWARP RNICs, because the remote receivers that do not have hardware iWARP can just use softiwarp.

An early version of softiwarp was developed at the Ohio Supercomputer Center (OSC) [61]. Last year IBM Research labs in Zurich, Switzerland, gave a presentation in which they demonstrated a running softiwarp implementation [62], and they have since released it as open-source software. This software runs as a module in the Linux kernel, utilizes the kernel's TCP sockets, and can be seen as a device driver

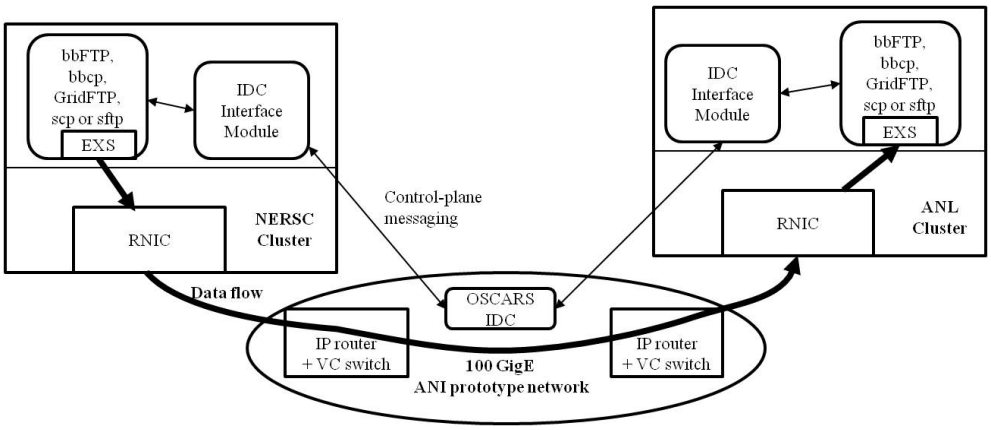

Figure 4. Experimental setup for testing the modified file transfer applications

by the OFED stack.

4.3.2 EXS integration into file transfer applications

Depending on the availability of open-source software and suitable support environments, we plan to modify two or more commonly used high-throughput secure file transfer applications, such as those listed in [63], which includes bbFTP [13], bbcp [64, 65], GridFTP [12], scp [66], and sftp [67]. Given the socket-based history of file transfer software, EXS should be an ideal API to quickly modify a user-space version of a file transfer client and server to use the RDMA networking technologies of RoCE and iWARP. Modified software will be tested within a local testbed at the University of New Hampshire, and managed in the NMI build and test facility.

4.3.3 Integration of circuit scheduling client with file transfer applications

An IDC Interface Module (IDCIM) has been implemented at the University of Virginia for a DOE-sponsored hybrid networking project [68], the purpose of which is to interface with a circuit scheduler (referred to as IDC in the OSCARS project) to request reservations, and the provisioning and release of circuits. This will modified for use in this project as needed. This module was developed starting with the the OSCARS Java Client API [19], provided by ESnet. The IDCIM does not need to be integrated directly into the file transfer applications, as it can be run externally using scripts before file transfers are initiated. The IDCIM implementation and testing will be done on a local testbed at UVA, and managed in the NMI build and test facility. ESnet provides a test IDC system [69] against which our software can be tested.

4.3.4 Experiments

Fig. 4 shows the configuration that will be used for the experiments. It shows that the modified file transfer applications with the EXS library, in concert with the IDC interface module, will be executed on the network nodes at the NERSC and ANL clusters. With the iWARP RNICs, the 100 GigE ANI network nodes will be configured as IP routers, and with the RoCE RNICs, Ethernet VLAN based virtual circuits will be established using the IDCIM and IDC.

For disk-to-disk file transfer across the wide-area network, all three datacenter networks, Fibre Channel, InfiniBand and Ethernet, will be involved. Fig. 2 shows that Fibre Channel is used between I/O Nodes and the disks. GPFS file systems will be mounted on the Network Nodes shown in Fig. 2. A GPFS system call will result in InfiniBand communication between the GFPS client on the Network nodes and GPFS servers on the I/O nodes. The file transfer application executing on the network nodes will then use iWARP or RoCEE to transfer the data over an IP-routed path or VLAN virtual circuit, respectively, between the two

datacenters.

The *objective* of these experiments is to demonstrate as close a throughput to 100 Gbps as possible for disk-to-disk large transfers. ANI will support advance circuit reservations, but since this test will require a full link capacity reservation, and other researchers will also be using ANI, we will request special permission to conduct this test during lean hours. Measurements taken by NERSC show a maximum throughput of 11 GB/sec (88 Gbps) from the I/O nodes, but this will be increased to 100 Gbps in time for the ANI deployment [70]. Between a pair of computers across the 4x QDR InfiniBand, a maximum rate of 3 GB/sec (24 Gbps) was measured. The 4x QDR configuration has a 40 Gbps signaling rate but can achieve a maximum of 32 Gbps given the 8B/10B encoding. Also, as the Mellanox RNICs are PCIe Gen 2 cards, and PCIe Gen 2 has a rate limit of 26 Gbps [71], this explains the measured value. However, since InfiniBand is a switch based communication fabric rather than a shared-medium one, this 24 Gbps rate can be simultaneously sustained between multiple pairs of computers. With GPFS, blocks of data from individual files can be striped across multiple disks, and as these blocks can be transferred by multiple I/O nodes, aggregate file transfer rates of 88-100 Gbps should be feasible.

4.4 Trials and technology transfer

As stated in Section Section 1.2.4, the goal of this step is to encourage and help scientists to adopt the networking solutions and modified applications developed in this project. We plan to start with the CESM scientists and have them test the modified applications in their regular work environments. The Earth System Grid [72–74] that provided a distribution method for data associated with the IPCC-AR4 data could be another likely adopter within the climate community.

For a more broad technology transfer of our work, the ESnet generated network requirements workshop reports will be mined for information on scientific projects that require large data set movement. These reports are available for the six Program areas within DOE's Office of Science: Basic Energy Sciences (BES) [75], Biological and Environmental Research (BER) [76], Nuclear Physics (NP) [77], Fusion Energy Sciences [78], High Energy Physics (HEP) [79], and ASCR [80]. Several of these teams will be approached for testing and adoption of our modified applications, and for implementation of our networking solutions.

5 Broader Impact

Integration of research and education The proposed project has many opportunities to integrate education. A course module on datacenter networking will be jointly developed by the UNH and UVA PIs, and offered as part of courses in the computing curriculum at both universities. We actively recruit top undergraduate students to undertake independent study courses through which these students participate in projects such as the proposed one. For example, the UVA PI guided two such students in Spring 2010, four in Fall 2010, and is currently guiding two students this semester (Spring 2011). This avenue will be used to engage our excellent undergraduate students in various research activities of the proposed project. The NCAR PI has mentored a number of students through the Summer Internships in Parallel Computational Science (SiParCS) program [81]. This is a 10-week program that engages undergraduate and graduate students in hands on experience in Computational science. The UNH PI has mentored numerous graduate and undergraduate students at the UNH Interoperability Laboratory (IOL) in projects associated with the lab's testing activities.

The Computer Science Department at UNH and the IOL actively recruit high school students considering admission to computer science at UNH to undertake part-time work at IOL during their entire undergraduate career starting freshman year. As freshmen and sophomores these undergrads are mentored by older students and staff, and as juniors and seniors they themselves become mentors for the new recruits. The practical applications of networking they experience working at IOL and the mentoring in which they

participate form an invaluable experience for undergraduates. We have budgeted funds to support one undergraduate for the life of this project, as well as funds to expand the joint CS Department/IOL outreach activities for high school students, for which the UNH PI will continue to serve as a faculty co-adviser.

Integrating diversity UVA's Engineering school has a Center for Diversity in Engineering [82]. The UVA PI has been an active supporter of programs organized by this center. For example, this PI participates in summer outreach programs, such as the 2009 Introduction to Engineering (ITE) program, in which about fifty high-school students participated, and has served as research adviser for undergraduate African American students in the Center's NSF Research Experience for Undergraduates program. One of these students has continued research under the PI's guidance in the past Fall and this Spring semesters with independent study courses. The PI also actively participates in the Center-sponsored SEAS Women Faculty & Graduate Student events.

In this proposed project, we have allocated funds under "participant costs," which will be used to sponsor programs such as the High School Visitation Weekend Program organized by the Society of Women Engineers (SWE) at UVA, which invites up to 100 female junior and senior high school students to spend a weekend exploring majors and careers in engineering. The PI will serve as a faculty adviser in organizing this program.

Impact on the NSF community Given the objectives of this project, its outcomes will have a broad impact on the NSF scientific community. As described in Section 6, the planned trials are targeted at moving our software from experimental to production user, and will target projects from six science areas. By choosing applications that are popular among scientists across many science areas (for modification to leverage the benefits of the new datacenter and wide-area networking technologies), we anticipate achieving a significant adoption rate.

6 Project timetable with deliverables

- **Year 1**:
  - Data analysis to gain an understanding of CESM projects' wide-area networking needs
    - Step 1: Plan and implement monitoring mechanisms such as GridFTP logging, tcpdumps, IPM output for MPI I/O, and PerfSONAR OWAMP for NCAR scientists: NCAR and UVA
    - Step 2: Analyze collected data: UVA
  - Intra-datacenter MPI applications: evaluate IB, RoCE and iWARP interconnects for a subset of CESM applications, NAS, NERSC-6 and HPCC benchmarks
    - Step 1: Execute applications and collect measurements: NCAR and UNH
    - Step 2: Analyze measurements collected from the above tests: UVA
    - Step 3: Write reports and publish papers: UVA, UNH and NCAR
  - Software implementation
    - Complete EXS library development and documentation: UNH
    - Integrate IDC interface module for virtual circuits with file transfer applications: UVA
  - Broader impact activities
    - Prepare course module on datacenter networking: UNH and UVA
    - Participate in diversity related activities: UVA, UNH and NCAR
- **Year 2**:
  - Data analysis to gain an understanding of CESM projects' wide-area networking needs (contd.)
    - Step 3: Deploy monitoring mechanisms in two or more CESM participating universities: NCAR and UVA

        - Step 4: Analyze collected measurements: UVA
        - Step 5: Write reports and publish papers: UVA and NCAR
    - NERSC-ANL MPI I/O experiments across ANI: evaluate RoCE/VLAN and iWARP/IP-routed paths for different MPI applications
        - Step 1: Execute applications and collect measurements: NCAR, UNH and UVA
        - Step 2: Analyze measurements: UVA
        - Step 3: Write reports and publish papers: UVA, UNH and NCAR
    - Software implementation
        - Integrate EXS library into a subset of file transfer applications: UNH
        - Test integrated file transfer applications: UVA
    - NERSC-ANL 100 GigE ANI file transfer experiments: evaluate networking solutions for file transfer applications
        - Step 1: Execute modified file transfer applications and collect measurements: UNH and UVA
        - Step 2: Analyze measurements: UVA and UNH
        - Step 3: Write reports and publish papers: UVA, UNH and NCAR
    - Broader impact activities
        - Teach course module on datacenter networking: UNH and UVA
        - Participate in diversity related activities: UVA, UNH and NCAR

- **Year 3**:
    - Trials and technology transfer
        - Step 1: Plan and arrange for the use of high-speed wide-area paths for CESM researchers between NCAR, NICS, SDSC, NERSC and ANL: UVA
        - Step 2: Execute and collect measurements for modified file transfers and wide-area MPI I/O applications: NCAR
        - Step 3: Analyze results and make modifications as needed for adoption: UVA, UNH and NCAR
        - Step 4: Transfer applications and networking solutions to other science projects: UNH
        - Step 5: Write reports and publish papers: NCAR, UVA and UNH
    - Broader impact activities
        - Teach course module on datacenter networking: UNH and UVA
        - Participate in diversity related activities: UVA, UNH and NCAR
        - Write report on activities with measures of success: UVA, UNH and NCAR

7 Compliance with requirements for SDCI proposals

This proposal meets the requirements specified in Section II.3 of the NSF SDCI solicitation as follows:

- Identification of the software focus area in the title of the proposal: *"SDCI Net:" included in title.*
- Direct engagement with one or more scientific research projects: *CESM project member is a PI.*
- Experimental deployment, trial use, or initial operational integration in a production environment supporting scientific research: *See Sections 1.2.4 and 4.4.*
- A pre-existing software base: *Initial versions of the EXS library (Section 4.3.1) and IDCIM (Section 4.3.3) already exist, and will serve as starting points for the software developed in this project. Similarly, existing software will be modified as necessary for our large file transfer work (Section 1.2.3).*

- Use of the NSF-funded NMI Build and test services: *See Section 4 and the Data Management Plan*.
- Identification of multiple application areas in science or engineering: *See Section 4.4*.
- The project plan must include milestones with release targets of software, deployment goals, and an evaluation plan: *See Sections 1.3 and 6. The evaluation plan is to use Sustained System Performance (SSP) for MPI applications (Section 1.2.2), and throughput for file transfers (Section 1.2.3). Ultimately, we recognize that the measure of success of this project will be in the adoption of our solutions by the scientific community. The participation of the CESM (NCAR) PI, and our plans for technology transfer as described in Section 4.4 are key to achieving this measure of success*.
- A compelling discussion of the software's potential use by broader communities: *See Section 1.1 for a description of CESM use cases, and more broadly, our understanding of the requirements of the six DOE Office of Science program area projects (Section 4.4)*.
- Identification of the open source license to be used: *Data Management Plan (GNU GPLv3 license)*.
- Collaborations with industry are encouraged where appropriate: *At present, we have none, but will seek partnerships with Mellanox for datacenter networking, and Juniper for wide-area networking, as the project proceeds. Also, we will advertise our EXS library to the Open Fabrics Alliance*.

8 Results from Prior NSF Support

Veeraraghavan: NSF 0335190, $2,113,000, Jan. 04-Dec. 08, End-To-End Provisioned Optical Network Testbed for Large-Scale eScience Applications: Four PhDs and six Masters degree students, two postdoctoral fellows, and two undergraduate research assistants were supported. The key research contributions consist of (i) a hybrid architecture for supporting both connectionless and circuit-switched dynamic services with wide-area testbed demonstrations; (ii) integration with file-transfer and remote-visualization applications for the TSI astrophysics project, demonstrating high-speed connectivity from NCSU to ORNL; (iii) an analytical characterization of immediate-request versus advance-reservation modes of bandwidth sharing; (iv) development of multiple aspects of file transfers over dynamically shared circuits, such as Circuit-TCP (CTCP) transport protocol and integration of a signaling client and CTCP into applications, such as web servers/proxies; (v) analytical and simulation models for resource sharing aspects of advance-reservation services; and (vi) secure control-plane design for GMPLS networks. Papers, design documents and software modules are available on the CHEETAH Web site [45].

Russell: Directed industry funded work with storage protocols (iSCSI) and Remote Direct Memory Access protocols (InfiniBand, iWARP) that included design, implementation and performance analysis of Linux kernel modules and user APIs [60]. See http://www.iol.unh.edu/services/testing /iwarp/training/rdr-exs-2.pdf.

Dennis: Co-investigator on NSF Award OCI-0749206, PetaApps: New Coupling Strategies & Capabilities for Petascale Climate Modeling" in the amount of $587,847 for the period of 01/01/08 to 12/31/11. In collaboration with the DOE Grand Challenge project, we have enabled the first ever century-long coupling of a 0.5 atmospheric and 0.1 ocean and sea ice components in a high-resolution climate simulation. To support the computing needs of this research project we were granted a TeraGrid allocation of 35M CPU hours at the National Institute for Computational Science for the proposal "The Role of Climate System Noise in Climate Simulations." This project regularly utilizes high-performance wide area networking to analyze the 120 TB of output data.

Woitaszek: Senior personnel on NSF-MRI Grant CNS-0821794, MRI-Consortium: Acquisition of a Supercomputer by the Front Range Computing Consortium (FRCC). This grant provided support for a consortium consisting of the University of Colorado at Boulder, University of Colorado at Denver, and the National Center for Atmospheric Research to acquire a 16,416-core, 184 TFLOP/s supercomputer. The system is anticipated to be in production in Q2 2011 serving researchers from the three institutions.

References

[1] L. Xu, K. Harfoush, and I. Rhee, "Binary increase congestion control for fast long-distance networks." in *Proceedings of IEEE INFOCOM*, Mar. 2003.

[2] S. Ha, I. Rhee, and L. Xu, "CUBIC: a new TCP-friendly high-speed TCP variant," *SIGOPS Oper. Syst. Rev.*, vol. 42, no. 5, pp. 64–74, 2008.

[3] S. Floyd, "HighSpeed TCP for large congestion windows," Feb. 2003. [Online]. Available: http://www.icir.org/floyd/hstcp.html

[4] C. Jin, D. X. Wei, and S. H. Low, "FAST TCP: Motivation, Architecture, Algorithms, Performance," in *Proceedings of IEEE INFOCOM*, Mar. 2004.

[5] T. Kelly, "Scalable TCP: Improving performance in highspeed wide area networks," Computer Communication Review 32(2), Apr. 2003.

[6] Y. Gu and R. L. Grossman, "SABUL: A transport protocol for grid computing." *Journal of Grid Computing*, 2003.

[7] Y. Gu, X. Hong, and R. L. Grossman, "Experiences in design and implementation of a high performance transport protocol," in *SC '04: Proceedings of the 2004 ACM/IEEE conference on Supercomputing*. Washington, DC, USA: IEEE Computer Society, 2004, p. 22.

[8] P. Datta, W. chun Feng, and S. Sharma, "End-system aware, rate-adaptive protocol for network transport in lambdagrid environments," SC06, Nov. 2006. [Online]. Available: http://sc06.supercomputing.org/schedule/pdf/pap229.pdf

[9] E. He, J. Leigh, O. Yu, and T. A. DeFanti, "Reliable Blast UDP: Predictable high performance bulk data transfer." in *Cluster Computing, 2002. Proceedings. 2002 IEEE International Conference on*, Sep. 2002, pp. 317–324.

[10] X. Zheng, A. P. Mudambi, and M. Veeraraghavan, "FRTP: Fixed rate transport protocol – a modified version of SABUL for end-to-end circuits," in *Proc. of Pathnets 2004 Workshop in IEEE Broadnets 2004*, Sept. 2004.

[11] A. P. Mudambi, X. Zheng, and M. Veeraraghavan, "A transport protocol for dedicated end-to-end circuits," in *Proc. of IEEE ICC 2006*, Istanbul, Turkey, Jun. 2006.

[12] GridFTP. [Online]. Available: http://dev.globus.org/wiki/GridFTP

[13] NASA research and engineering network bbFTP. [Online]. Available: http://www.nren.nasa.gov/bbftp.html

[14] Tom Talpey and Paul Grun, "Remote Direct Memory Access over the Converged Enhanced Ethernet Fabric: Evaluating the Options," in *Proc. of IEEE HOT Interconnects*, 2009. [Online]. Available: http://www.hoti.org/hoti17/program/slides/Panel/Talpey_HotI_RoCEE.pdf

[15] Neteffect, Understanding iWARP: Eliminating Overhead and Latency in multi-Gb Ethernet Networks. [Online]. Available: http://download.intel.com/support/network/adapter/pro100/sb/understanding_iwarp.pdf

[16] D. Skinner, "Integrated performance monitoring: A portable profiling infrastructure for parallel applications," in *Proc. ISC2005: International Supercomputing Conference*, Heidelberg, Germany, 2005.

[17] "Performance analysis of high performance computing applications on the Amazon web services cloud," in *IEEE Cloudcom*, 2010.

[18] Documents and information from past workshops on science requirements for esnet networking. [Online]. Available: http://www.es.net/hypertext/requirements.html

[19] OSCARS Java Client API. [Online]. Available: https://wiki.internet2.edu/confluence/display/DCNSS/Java+Client+API

[20] R. Recio, B. Metzler, P. Culley, J. Hilland, and D. Garcia, "A Remote Direct Memory Access Protocol Specification," IETF RFC 5040, Oct. 2007.

[21] InfiniBand Trade Association. (2007, Nov.) InfiniBand Architecture Specification Volume 1, Release 1.2.1. [Online]. Available: http://infinibandta.org

[22] H. Shah, J. Pinkerton, R. Recio, and P. Culley, "Direct Data Placement over Reliable Transports," IETF RFC 5041, Oct. 2007.

[23] P. Culley, U. Elzur, R. Recio, S. Bailey, and J. Carrier, "Marker PDU Aligned Framing for TCP Specification," IETF RFC 5044, Oct. 2007.

[24] P802.1Qbb: Priority-based Flow Control. [Online]. Available: http://www.ieee802.org/1/files/public/docs2008/bb-pelissier-pfc-proposal-0508.pdf

[25] P802.1Qaz: Enhanced Transmission Selection (aka Priority Groups). [Online]. Available: http://www.ieee802.org/1/files/public/docs2008/az-wadekar-ets-proposal-0608-v1.01.pdf

[26] P802.1Qaz: DCB Capability Exchange Protocol (DCBX). [Online]. Available: http://www.ieee802.org/1/files/public/docs2008/az-wadekar-dcbx-capability-exchange-discoveryprotocol-1108-v1.01.pdf

[27] OpenFabrics Alliance, "http://www.openfabrics.org," 2009.

[28] OpenFabrics Enterprise Distribution, "http://www.openfabrics.org/," 2009.

[29] Open-iSCSI, "http://www.open-iscsi.org," 2007.

[30] E. Burns and R. Russell, "Implementation and Evaluation of iSCSI over RDMA," in *Proceedings of the 5th International Workshop on Storage Network Architecture and Parallel I/O (SNAPI'08)*, Sep. 2008.

[31] OpenFabrics Alliance, "OpenFabrics iSER," 2009, http://wiki.openfabrics.org/tiki_index.php?page=iSER.

[32] M. Patel and R. Russell, "Design and Implementation of iSCSI Extensions for RDMA," Sep. 2005, University of New Hampshire InterOperability Laboratory.

[33] Open MPI: Open Source High Performance Computing, "http://www.open-mpi.org," 2009.

[34] MVAPICH: MPI over InfiniBand and iWARP, "http://mvapich.cse.ohio-state.edu," 2009.

[35] Interconnect Software Consortium in association with the Open Group, "Extended Sockets API (ES-API) Issue 1.0," Jan. 2005.

[36] Open Group, "http://www.opengroup.org," 2009.

[37] Extended Sockets API (ES-API), Issue 1.0. [Online]. Available: http://www.opengroup.org/pubs/catalog/c050.htm

[38] L. Guo and I. Matta, "The war between mice and elephants," in *Network Protocols, 2001. Ninth International Conference on*, 2001, pp. 180 – 188.

[39] On-demand Secure Circuits and Advance Reservation System (OSCARS). [Online]. Available: https://oscars.es.net/OSCARS/

[40] Internet2 ION Circuit Reservation Demo. [Online]. Available: http://iondemo.net.internet2.edu:8080/ion/

[41] Bandwidth on Demand with AutoBAHN. [Online]. Available: http://www.geant2.net/server/show/ConWebDoc.2544

[42] Canarie's user-controlled lightpaths (UCLP). [Online]. Available: http://www.canarie.ca/canet4/uclp/index.html

[43] DICE InterDomain Controller Protocol (IDCP). [Online]. Available: http://hpn.east.isi.edu/dice-idcp/

[44] TeraPaths: Configuring End-to-End Virtual Network Paths with QoS Guarantees. [Online]. Available: https://www.racf.bnl.gov/terapaths/

[45] CHEETAH. [Online]. Available: http://cheetah.cs.virginia.edu

[46] The Lambda Station Project. [Online]. Available: http://www.lambdastation.org/

[47] Community Earth System Model (CESM). [Online]. Available: http://www.cesm.ucar.edu/

[48] M. Parry, O. Canziani, J. Palutikof, P. van der Linden, C. Hanson, and et al., "Climate change 2007: Impacts, adaptation and vulnerability," 2007, Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, 976 pages. [Online]. Available: http://www.ipcc.ch/ipccreports/ar4-wg2.htm

[49] Project Athena: High Resolution Global Climate Simulations. [Online]. Available: http://wxmaps.org/athena/home/

[50] B. Draney. Magellan: NERSC Cloud Testbed. [Online]. Available: http://www.nersc.gov/nusers/systems/magellan/

[51] LBNL Advanced Network Initiative Design Document. [Online]. Available: http://sc.doe.gov/ascr/ProgramDocuments/ARRA/ANIDesignDoc.pdf

[52] B. Draney. Magellan. [Online]. Available: http://indico.fnal.gov/materialDisplay.py ?contribId=6&materialId=slides&confId=2970

[53] PerfSONAR One-Way Ping (OWAMP) Version 3.1. [Online]. Available: http://www.internet2.edu/performance/owamp/

[54] D. H. Bailey, E. Barszcz, J. T. Barton, D. S. Browning, R. L. Carter, R. A. Fatoohi, P. O. Frederickson, T. A. Lasinski, H. D. Simon, V. Venkatakrishnan, and S. K. Weeratunga, "The NAS parallel benchmarks," The International Journal of Supercomputer Applications, Tech. Rep., 1991.

[55] NERSC-6 application benchmarks. [Online]. Available: http://www.nersc.gov/projects/SDSA/software/?benchmark=NERSC6

[56] P. Luszczek, J. J. Dongarra, D. Koester, R. Rabenseifner, B. Lucas, J. Kepner, J. Mccalpin, D. Bailey, and D. Takahashi, "Introduction to the HPC challenge benchmark suite," http://icl.cs.utk.edu/projectsfiles/hpcc/pubs/ hpcc-challenge-benchmark05.pdf, 2005.

[57] P. Colella, "Defining software requirements for scientific computing," 2004.

[58] K. Asanovic, R. Bodik, B. Catanzaro, J. Gebis, P. Husbands, K. Keutzer, D. Patterson, W. Plishker, J. Shalf, S. Williams, and K. Yelick, "The landscape of parallel computing research: A view from Berkeley," Electrical Engineering and Computer Sciences, University of California at Berkeley, Tech. Rep. UCB/EECS-2006-183, December 2006.

[59] R. Russell, "The Extended Sockets Interface for Accessing RDMA Hardware," in *Proceedings of the 20th IASTED International Conference on Parallel and Distributed Computing and Systems (PDCS 2008)*, T. Gonzalez, Ed., Nov. 2008, pp. 279–284.

[60] ——, "A general-purpose API for iWARP and InfiniBand," in *Proc. of DC CAVES workshop*, September 2009. [Online]. Available: http://www.iol.unh.edu/services/testing/iwarp/training/rdr-exs-2.pdf

[61] D. Delessandro, A. Devulapalli, and P. Wyckoff, "Design and Implementation of the iWARP Protocol in Software." PDCS, Nov. 2005.

[62] B. Metzler, F. Neeser, and P. Frey, "A Software iWARP Driver for OpenFabrics," in *http:/www.openfabrics.org/archives/sonoma2009/monday/softiwarp.pdf*, Mar. 2009.

[63] File transfers at NERSC. [Online]. Available: http://www.nersc.gov/nusers/help/access/secureftp.php

[64] BBCP: P2P Data Copy Program. [Online]. Available: http://www.slac.stanford.edu/ abh/bbcp/

[65] BBCP. [Online]. Available: http://www.nccs.gov/user-support/general-support/data-transfer/bbcp/

[66] How the SCP protocol works. [Online]. Available: http://blogs.sun.com/janp/entry/how_the_scp_protocol_works

[67] FileZilla: the free FTP solution. [Online]. Available: http://sourceforge.net/projects/filezilla/

[68] Hybrid Network Traffic Engineering System (HNTES). [Online]. Available: http://www.ece.virginia.edu/mv/research/DOE09/documents/documents.html

[69] OSCARS test IDC system. [Online]. Available: https://test-idc.internet2.edu:8443/axis2/services/OSCARS

[70] "Conversations with Brent Draney, Magellan project, NERSC," Apr. 2010.

[71] A Margalla Communications Special Report. High-speed Remote Direct Memory Access (RDMA) Networking for HPC; Comparative Review of 10 GbE iWARP and InfiniBand. [Online]. Available: http://margallacomm.com/downloads/MicrosoftWord_rdma_wp_0210.pdf

[72] D. N. Williams, D. E. Bernholdt, I. T. Foster, and D. E. Middleton, "The earth system grid center for enabling technologies: Enabling community access to petascale climate datasets," *CTWatch Quarterly*, vol. 3, no. 4, November 2007.

[73] D. N. Williams, R. Drach, R. Ananthakrishnan, I. T. Foster, D. Fraser, F. Siebenlist, D. B. Berndoldt, M. Chen, J. Schwidder, S. Bharathi, A. L. Cherzenak, R. Shulder, M. Su, D. Brown, L. Cinquini, P. Fox, J. Garcia, D. E. Middleton, W. G. Strand, N. Wilhelmi, S. Hankin, R. Schweitzer, P. Jones, A. shoshani, and A. Sim, "The earth system grid: Enabling access to multimodel climate simulation data," *Bull. Amer. Meteor. Soc.*, vol. 90, pp. 105–205, February 2009.

[74] Earth system grid center for enabling technologies (ES-CET) – ESG. [Online]. Available: http://esg-pcmdi.llnl.gov/

[75] BES Science Network Requirements: Report of the Basic Energy Sciences Network Requirements Workshop Conducted June 4-5, 2007. [Online]. Available: http://www.es.net/pub/esnet-doc/BES-Net-Req-Workshop-2007-Final-Report.pdf

[76] BER Science Network Requirements: Report of the Biological and Environmental Research Network Requirements Workshop Conducted July 26 and 27, 2007. [Online]. Available: http://www.es.net/pub/esnet-doc/BER-Net-Req-Workshop-2007-Final-Report.pdf

[77] NP Science Network Requirements: Report of the Nuclear Physics Network Requirements Workshop Conducted May 6 and 7, 2008. [Online]. Available: http://www.es.net/pub/esnet-doc/NP-Net-Req-Workshop-2008-Final-Report.pdf

[78] FES Science Network Requirements: Report of the Fusion Energy Sciences Network Requirements Workshop Conducted March 13 and 14, 2008. [Online]. Available: http://www.es.net/pub/esnet-doc/FES-Net-Req-Workshop-2008-Final-Report.pdf

[79] HEP Science Network Requirements: Office of High Energy Physics Network Requirements Workshop Conducted August 27 and 28, 2009 DRAFT Report. [Online]. Available: http://workshops.es.net/2009/hep-net-req/wiki/pub/HEPNetReq/DraftReportPDF/HEP-Net-Req-Workshop-2009-DRAFT-Final-Report-v3.pdf

[80] ASCR Science Network Requirements: Office of Advanced Scientific Computing Research Network Requirements Workshop Conducted April 15 and 16, 2009, Final Report. [Online]. Available: http://www.es.net/pub/esnet-doc/ASCR-Net-Req-Workshop-2009-Final-Report.pdf

[81] Summer interships in parallel computational science. [Online]. Available: http://www2.cisl.ucar.edu/siparcs

[82] Center for Diversity in Engineering at the University of Virginia. [Online]. Available: http://www.seas.virginia.edu/admin/diversity/